JOURNAL OF APPLIED STATISTICS
2021, VOL. 48, NO. 6, 1091–1110
https://doi.org/10.1080/02664763.2020.1757048

Taylor & Francis
Taylor & Francis Group

A signature enrichment design with Bayesian adaptive randomization

Fang Xia[a], Stephen L. George[b], Jing Ning[a], Liang Li[a] and Xuelin Huang[a]

[a]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA;
[b]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

ABSTRACT
Clinical trials in the era of precision cancer medicine aim to identify and validate biomarker signatures which can guide the assignment of individually optimal treatments to patients. In this article, we propose a group sequential randomized phase II design, which updates the biomarker signature as the trial goes on, utilizes enrichment strategies for patient selection, and uses Bayesian response-adaptive randomization for treatment assignment. To evaluate the performance of the new design, in addition to the commonly considered criteria of Type I error and power, we propose four new criteria measuring the benefits and losses for individuals both inside and outside of the clinical trial. Compared with designs with equal randomization, the proposed design gives trial participants a better chance to receive their personalized optimal treatments and thus results in a higher response rate on the trial. This design increases the chance to discover a successful new drug by an adaptive enrichment strategy, i.e. identification and selective enrollment of a subset of patients who are sensitive to the experimental therapies. Simulation studies demonstrate these advantages of the proposed design. It is illustrated by an example based on an actual clinical trial in non-small-cell lung cancer.

ARTICLE HISTORY
Received 24 February 2019
Accepted 11 April 2020

KEYWORDS
Adaptive design; precision medicine; biomarker; Bayesian adaptive randomization; classifier; enrichment strategy

2010 MATHEMATICS SUBJECT CLASSIFICATION
62P10

1. Introduction

Most human cancers are heterogeneous with respect to their molecular, genomic and phenotypic properties and treatments that target specific molecules may only benefit a subset of patients [15,26]. Thus, it is important to design efficient trials that offer optimal treatments for as many patients as possible based on their biomarker profiles.

In contrast to traditional non-adaptive design methods, adaptive designs have recently gained popularity due to their flexibility and efficiency. In 2006, the Pharmaceutical Research and Manufactures of America (PhRMA) Working Group defined an adaptive design as a design that allows modification of the on-going study based on accumulating data, without undermining the validity and integrity of the trial [10]. Examples of adaptive designs include those that use adaptive randomization [18], sample size re-estimation [16],

CONTACT Xuelin Huang ✉ xlhuang@mdanderson.org

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02664763.2020.1757048

changes to eligibility criteria [41], and early stopping rules for safety, futility or efficacy [23]. Adaptive randomization procedures can use a frequentist framework [19] or a Bayesian framework [5]. Commonly used adaptive randomization methods include treatment-adaptive randomization, covariate adjustment randomization and response adaptive randomization [43]. Bayesian adaptive randomization methods allow the combination of prior knowledge and observed data to learn about parameters of interest [11]. Bayesian adaptive randomization procedures apply Bayes theorem repetitively based on accumulating data to adjust the randomization probability for each newly enrolled patient. Such procedures may be considered more ethical by assigning more patients to the more effective treatment arm. Several randomized phase II clinical trials have adopted Bayesian adaptive randomization, including BATTLE and I-SPY2 [2,14,22,36]. Randomized phase II trials are becoming attractive in the modern era because they greatly enhance the potential for biomarker discovery and can assure optimal use of limited phase III financial and patient resources [24]. Based on this consideration, in this article, we propose a new design for randomized phase II trials, which are exploratory rather than confirmatory in nature.

Recently, the rapid advancement of biomarker studies in oncology has promoted the development and application of precision medicine, previously known as 'personalized medicine' [12,25]. Precision medicine targets a subpopulation of patients who are most likely to respond to the treatment based on their characteristics or biomarker profile. Prognostic biomarkers provide information on clinical outcome independently of the treatment received, while predictive biomarkers provide information on clinical outcome for a particular treatment [31]. Prognostic biomarkers can be used to screen good and bad prognostic patients, while predictive biomarkers can be used to measure how likely the patient will respond to a particular treatment. Before using biomarker information in clinical practice, it is essential to test the biomarkers for analytical and clinical validity and clinical utility [33]. For the designs described in this paper, we focus on predictive biomarkers.

A biomarker adaptive design utilizes patient biomarker information measured at baseline to allow adaptation based on previous patients' responses. When the biomarker classification used to categorize the sensitive status for potential patients are known before the start of the clinical trial, a biomarker adaptive design is particularly useful. If the biomarker classification is unknown, one may consider an adaptive signature design [9] or the cross-validated variation of this design [7]. These designs use an equal randomization approach but include a development stage and a validation stage to determine a sensitive subset classifier. An adaptive enrichment design [35] allows changes to the enrollment criteria as the trial proceeds. Restricting enrollment to patients who are most likely to benefit from the treatment based on a classifier improves the efficiency of the clinical trial but may require a longer time to accrue the targeted patients.

In this article, we propose a signature enrichment design with Bayesian response-adaptive randomization (SEDAR). This design combines elements of adaptive signature designs and adaptive enrichment designs. During patient enrollment, SEDAR uses an enrichment strategy which oversamples the sensitive patient cohort and undersamples the non-sensitive patients to form a selected mixture of the patient population. It also applies Bayesian response-adaptive randomization, adaptively adjusting future allocation probabilities based on the accumulated data. As a result, SEDAR has the advantages of Bayesian adaptive randomization as well as an enrichment strategy, yielding a higher overall response rate on the trial and other advantages described below. At the end of the trial, we test the treatment effect (1) in the selected mixture of sensitive and non-sensitive patients and (2) in the sensitive patient subgroup, with the control of overall Type I error rates.

In addition to the usual evaluation of statistical power, we propose four trial evaluation criteria to assess the performance of the designs: current individual loss (CIL), future individual loss (FIL), the probability of an individual in the trial receiving optimal treatment (PCO) and the probability of a future individual receiving personalized optimal treatment (PFO). These criteria evaluate the benefits for both trial participants and future patients from the perspective of precision medicine, providing a more comprehensive assessment of trial performance. Results are demonstrated in extensive simulations.

The rest of the paper is organized as follows. In Section 2, we briefly review the cross-validated adaptive signature design, the adaptive enrichment design, and our Bayesian response-adaptive randomization scheme. In Section 3, we describe our cross-validated signature enrichment design with response-adaptive randomization. In Section 4, we define the trial performance evaluation criteria. In Section 5, we report the results of simulation studies to compare our proposed design with the cross-validated adaptive signature design and the adaptive enrichment design. In Section 6, we present an example based on the Iressa Pan-Asia Study (IPASS) [28]. In Section 7, we conclude with a brief discussion.

2. Review of adaptive signature designs, adaptive enrichment designs and Bayesian adaptive randomization

As the name suggests, SEDAR combines cross-validated signature enrichment with Bayesian response-adaptive randomization. In this section, we briefly review the cross-validated adaptive signature design, the adaptive enrichment design, and the Bayesian response-adaptive randomization method. For all the designs mentioned in this article, we consider a two-arm clinical trial to compare the efficacy of an experimental treatment E and a control treatment C.

2.1. Adaptive signature design (ASD)

When a reliable classifier for identifying sensitive patients is not available at the start of the trial, an adaptive signature design [7,9] may be useful. If the overall test for treatment effect is not significant, this design defines a development stage and a validation stage for a sensitive patient subpopulation classifier. Since only a portion of patients contributes to each stage, the cross-validated extension of the adaptive signature design [7] adopts a K-fold cross-validation method to classify patients from the entire population and to adjust the P value using a permutation test [34].

This design uses equal randomization. If the overall test is not significant, it will divide patients into two cohorts: a development cohort and a validation cohort for determining a sensitive subset classifier. The design controls the overall Type I error rate at α, where $\alpha = \alpha_1 + \alpha_2$. If the overall treatment effect is statistically significant at some pre-specified level α_1, then the experimental treatment is considered globally beneficial. Otherwise, the adaptive signature design develops a sensitive patient classifier and tests whether the classifier is useful for treatment selection at Type I error rate α_2 (see Figure 1).

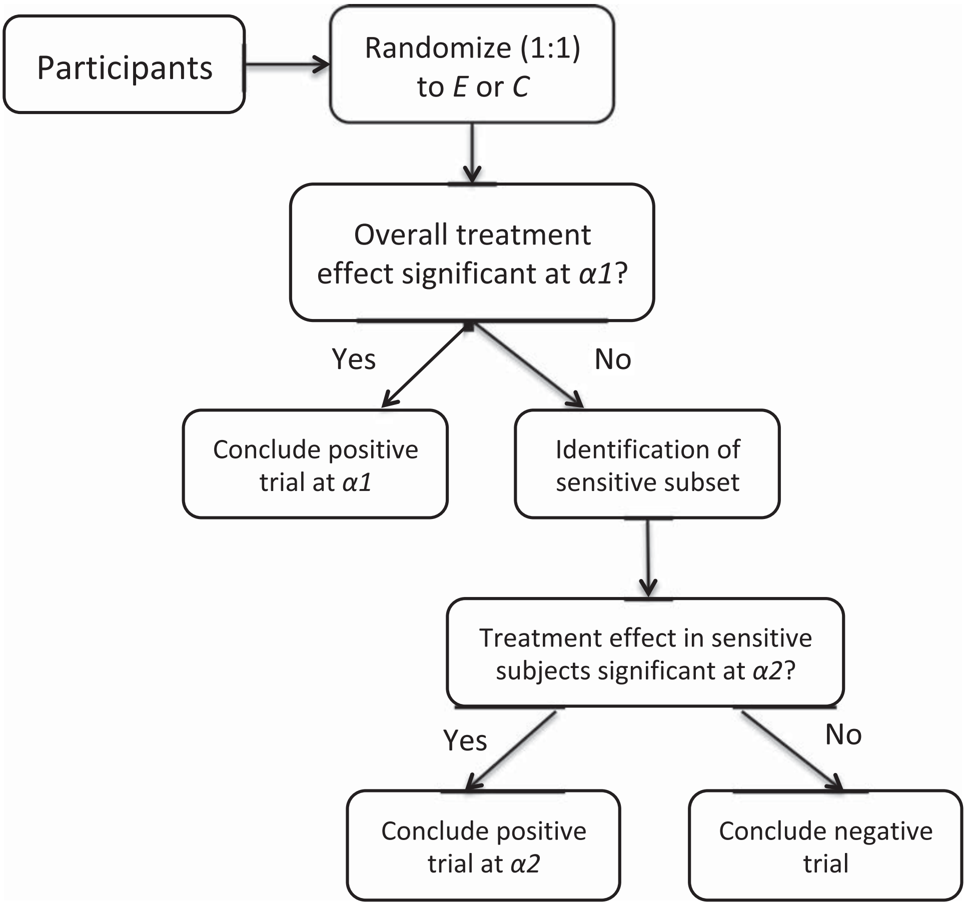
Participants
Randomize (1:1) to E or C
Overall treatment effect significant at $\alpha 1$?
Yes
No
Conclude positive trial at $\alpha 1$
Identification of sensitive subset
Treatment effect in sensitive subjects significant at $\alpha 2$?
Yes
No
Conclude positive trial at $\alpha 2$
Conclude negative trial

Figure 1. Diagram for adaptive signature design (ASD) [7,9].

The determination of a sensitive patient subgroup is based on machine learning voting methods in two stages. The first stage identifies predictive biomarkers. The second stage identifies sensitive patients based on the chosen predictive biomarkers from Stage I. A patient in the validation set is classified as sensitive if the predicted odds ratio of experimental arm versus standard arm exceeds some prespecified value. The treatment effect only in the sensitive patients in the validation cohort is then assessed at a prespecified reduced Type I error rate α_2 to test whether the classifier is useful for treatment selection.

The cross-validated adaptive signature design utilizes K-fold cross-validation to obtain a final sensitive patient subset from the entire trial population. Since the sensitive patient subset is obtained using cross-validation, a permutation test may be used to adjust the P value [7].

2.2. Adaptive enrichment design (AED)

The adaptive enrichment design [35] restricts enrollment to sensitive patients, as defined by an adaptive classifier function, and excludes all nonsensitive patients. Patients are

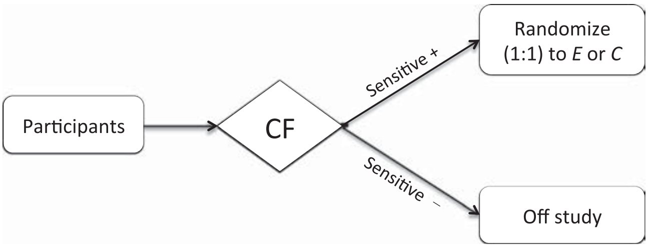

Figure 2. Diagram for adaptive enrichment design (AED) [35]. *CF* denotes the sensitive status classifier function.

grouped into blocks according to their entry time. For the initial patient block, equal randomization is used to allocate patients to treatment arms. For subsequent patient blocks, as illustrated in Figure 2, the probability of response is estimated for patient i with each treatment E and C, based on the cumulative estimated response rates from patients in the previous blocks. Let $\hat{r}_{i,x,E}$ and $\hat{r}_{i,x,C}$ denote the response rate of patient i with signatures x when receiving E or C, respectively. The classifier function determines the sensitive status, which is defined as

$$CF(\hat{r}_{i,x,E}, \hat{r}_{i,x,C}) = I(\hat{r}_{i,x,E} - \hat{r}_{i,x,C} > \delta), \tag{1}$$

where δ denotes the minimal clinical meaningful difference. Let $\hat{CF}_i$ refer to the latest estimate of the classifier function at the time of the entry of the ith patient. The classifier function restricts entry into the trial to patients with $\hat{CF}_i = 1$, the sensitive patients, and excludes patients whose $\hat{CF}_i = 0$, the non-sensitive patients. Newly enrolled patients will be assigned to one of the treatment arms using equal randomization. The classifier will be updated once a new block of data becomes available. The process is repeated until the desired total number of patients is reached.

As the adaptive enrichment design will include mostly sensitive patients (except in the initial patient block), it can improve the efficiency of the clinical trial and protect non-sensitive patients from exposure to ineffective treatments. However, it can be difficult to choose a strategy to update enrollment criteria and requires a longer time to accrue sufficient number of sensitive patients than the traditional all-comers designs. Also, if the treatment is effective in all patients, restricting enrollment may exclude patients who would benefit from participating in the trial [20].

2.3. Bayesian adaptive randomization

Adaptive randomization updates allocation probabilities based on currently enrolled patient responses, assigning more patients to the more effective treatments. Compared with equal randomization, response-adaptive randomization places more emphasis on individual benefit and advantages, especially with complicated designs [1]. Some objections to adaptive randomization include potential bias due to inhomogeneous patient population throughout the trial, statistical inefficiency due to unbalanced patient allocation and the requirement that responses be collected shortly after treatment [20]. Generally, the advantages of response-adaptive randomization are more evident with larger sample sizes and with larger differences between treatments.

Bayesian adaptive randomization assumes that the allocation probability for treatment E is proportional to the posterior probability $\hat{P}(r(i,E) > r(i,C)|data)$. The probability of patient i receiving treatment E, $P_{i,E}$, is given by [30,37] as

$$P_{i,E} = \frac{\hat{P}(r(i,E) > r(i,C)|data)^{w}}{\hat{P}(r(i,E) > r(i,C)|data)^{w} + \hat{P}(r(i,E) \leq r(i,C)|data)^{w}} \tag{2}$$

where w is a tuning parameter between 0 and 1. When w is equal to 0, there is equal randomization. When w is equal to 1, $P_{i,E}$ is the same as $\hat{P}(r(i,E) > r(i,C)|data)$. Common practice is to set w equal to 1/2 or to use $n/2N$ where n is the cumulative sample size [37].

3. Singature enrichment design with adaptive randomization (SEDAR)

We propose a cross-validated signature enrichment design with response-adaptive randomization. For the initial block, we use equal randomization and let the initial block size be twice that of the subsequent adaptive randomization blocks [42]. The framework for subsequent patient blocks of the design is presented in Figure 3. We assume patient responses can be observed shortly after receiving treatments. The enrollment procedure for future patients is as follows:

(1) For each patient, apply the current sensitive patients classifier function (CF) derived on data from previous blocks. This separates patients into one of the two categories: sensitive and non-sensitive.
(2) If the patient is determined to be sensitive, randomize to E or C based on the Bayesian adaptive randomization scheme.
(3) If the patient is determined to be non-sensitive, calculate the enrollment function (EF $= 1$ with probability P; EF $= 0$ with probability 1-P). If EF $= 1$, randomize based on the Bayesian randomization scheme as above; if EF $= 0$, the patient is off study (i.e. not randomized). The details of determining the enrollment probability are given in Section 3.1.
(4) After the current block is completely accrued, refine the classifier function based on the additional results from the current block and repeat the steps above for the next block.

As is common with all enrichment designs, it is difficult to calculate the required sample size a priori. To estimate the total required sample size for our design, we first calculate the sufficient sample size to reach the desired power using traditional methods for non-enrichment designs [13]. Then we use simulations to modify the sample size for the given design parameters.

3.1. Sensitive patient classifier function and enrollment function

After each block of data becomes available, we develop a sensitive patient classifier to screen for future patients. Many classification algorithms have been studied for biomarker classifier development [29,32]. Some popular choices include logistic regression, random forest [3], support vector machines [39] and diagonal linear discriminant analysis [6]. To estimate

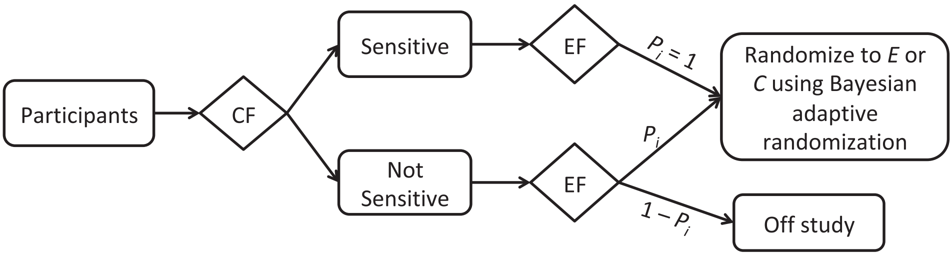

Figure 3. Diagram for signature enrichment design with adaptive randomization (SEDAR). *CF* is the classifier for determining patient sensitive status; *EF* is the enrollment function; P_i is the probability of enrolling the *i*th patient.

the probabilities of response for a patient with the experimental or control treatment, many statistical models or algorithmic methods, e.g. logistic regression and random forests, may be used [21]. For our proposed design, we use a random forest approach. In practice, a different classification algorithm may be used depending on trial-specific consideration. A brief description of the random forest technique is summarized below. More details can be found elsewhere [17].

Random forests grow a large collection of de-correlated decision trees [4]. Each tree starts with drawing a bootstrap sample from the original dataset. At each node, a best split is determined among a randomly selected subset of the total input variables. When a random forests approach is used for classification purpose, the class prediction is the majority vote from the total number of decision trees. The class probabilities are the proportion of votes from the ensemble of trees.

At an early stage of the clinical trial, the available patient data to build a classifier is limited. In contrast to the adaptive enrichment designs that restrict entry into the trial to sensitive patients, we use an enrichment strategy which will include both sensitive and non-sensitive patients but with different enrollment probabilities. Let δ denote the minimal clinical meaningful difference and $\hat{r}_{i,x,E}$ and $\hat{r}_{i,x,C}$ denote the estimated response rates for potential patient i with signatures x if receiving treatment E or C, respectively. The sensitive patient status is determined in the same way as AED, which is:

$$CF(\hat{r}_{i,x,E}, \hat{r}_{i,x,C}) = I(\hat{r}_{i,x,E} - \hat{r}_{i,x,C} > \delta), \tag{3}$$

Let $\hat{CF}_i$ refer to the latest estimate of the classifier function at the time of entry of the ith patient. If $\hat{CF}_i = 1$, then patient i is sensitive; otherwise, the patient i is non-sensitive. For sensitive patients, the probability of enrollment is 1. For non-sensitive patients, we use a continuous function as the enrollment function to determine the enrollment probability for patient i with signatures x. The enrollment probabilities are calculated as

$$P_i = \min\left\{\max\left(\frac{\hat{r}_{\{i,x,E\}} - \hat{r}_{\{i,x,C\}}}{\delta}, 0.10\right), 1\right\} \tag{4}$$

Thus, the probability for sensitive patients is 1 (i.e., all sensitive patients are enrolled) and for non-sensitive patients the enrollment probability is between 0.1 and 1. When all patient data are available, we build a final sensitive patient subset classifier. In our design, we use random forests as the algorithm for sensitive status classification.

During the trial process, we screen sensitive patients to enroll in the next block. This sensitive patient classifier is built on the previous patient response, treatment received and biomarker information. At the final analysis, the classifier is developed by cross-validation based on complete patient data, as described in the next subsection. The sensitive patient subset selected is used to test the treatment effect difference in the sensitive subgroup.

3.2. Bayesian decision rule

The proposed design tests the difference of the treatment effect between E and C (1) in the selected mixture of sensitive and nonsensitive patients and (2) in the sensitive patient subset using Bayesian decision rules, as described below.

To test the treatment effect in the sensitive population, we first identify the overall sensitive patient subset from the entire trial population using 10-fold cross-validation. This technique is also used in the cross-validated extension of the adaptive signature design [7]. For K-fold cross-validation, the entire trial population is randomly partitioned into K nonoverlapping equal-sized subpopulations. In the final analysis, one of the K folds is used as the validation cohort, and the rest $K-1$ folds are treated as the development cohort and a sensitive patient subset S_k $(k = 1, \ldots, K)$ is developed. This procedure is repeated K times. Each patient appears once in the validation cohort and $K-1$ times in the development cohort. Then the sensitive patient subset from the entire trial patients is $S = \bigcup_{k=1}^{K} S_k$.

At final analysis, arm E is claim efficacious if

$$\hat{Pr}(r_E > r_c \mid data) > a_U \tag{5}$$

where a_U is a cut-off parameter whose value is tuned to control the Type I error rates. The Type I error rate for testing treatment effect in the selected mixture of sensitive and nonsensitive patients is α_1, and for testing treatment effect in the sensitive subset it is α_2, so that the overall Type I error rate $\alpha = \alpha_1 + \alpha_2$. That is, we calibrate different a_U values for the selected mixture analysis and for the sensitive subset analysis to control their Type I error rates.

3.3. Required design parameters

The required design parameters (Table 1) are the sample size for the initial equal randomization block (N_0) and for the adaptive randomization blocks (N_{bar}). Then the total sample size $N = N_0 + m * N_{bar}$, where $m + 1$ is the total number of blocks. Additionally, the proposed design requires one to specify the tuning parameter w; minimal clinical difference δ for sensitive status classification; the Type I error rate for testing the treatment effect in the selected mixture of patients (α_1) and the Type I error rate for testing treatment effect in the sensitive subset only (α_2).

4. Trial performance evaluation criteria for simulations

In addition to commonly used operating characteristics such as the average response rate and power, we propose the following trial performance evaluation criteria for simulation studies of a signature clinical trial:

Table 1. Design parameters required for the signature enrichment design with Bayesian adaptive randomization.

N_0	Sample size for the initial equal randomization block
N_{bar}	Sample size for adaptive randomization blocks
N	Total sample size
w	Tuning parameter
δ	minimal clinical difference for sensitive status classification
α_1	Type I error rate for testing treatment effect in selected mixture of patients
α_2	Type I error rate for testing treatment effect in sensitive subset

- Current individual loss (CIL)
- Future individual loss (FIL)
- Probability of a current individual in the trial receiving personalized optimal treatment (PCO)
- Probability of a future individual receiving personalized optimal treatment (PFO)

For CIL and FIL, we first define a *match* for enrolled and future patients. For currently enrolled patients, a *match* occurs when the patient's actual treatment received is the same as the best treatment from the true model. For future patients, a *match* occurs when the superior treatment selected based on the final fitted model is the same as the best treatment from the true model. For a currently enrolled patient i with signature x, let $\hat{P}_i(Y=1|T,x)$ denote the probability of responding to the received treatment T. Similarly, for a potential future patient, $\hat{P}_i(Y=1|T,x)$ refers to the probability of responding to treatment T as determined by the fitted model. Let $\hat{P}_i(Y=1|T=OPT,x)$ denote the probability of responding to the *optimal* treatment determined by the true model. Then we define the personalized loss function as the follows:

$$Loss(i) = \begin{cases} 0 & \text{if it is a match} \\ \hat{P}_i(Y=1|T,x) - \hat{P}_i(Y=1|T=OPT,x) & \text{if there is not a match} \end{cases} \tag{6}$$

Then CIL and FIL are the average values of the individual loss for trial participants and for future patients, respectively. A small value of CIL indicates that most currently enrolled patients have received the treatment that they will respond at least similarly as the optimal treatments according to the true model. The interpretation is similar for FIL.

The PCO and PFO quantities are the probabilities of receiving the personalized optimal treatment for trial participants and future patients, respectively. PCO can be estimated as the average fraction of matches between the optimal treatment selection by the true model and by the patients' actual treatment received.

PFO is the probability of a future patient receiving the personalized optimal treatment if the treatment recommendation from the trial were followed. It can be estimated as the average fraction of matches for prospective patients between the optimal treatment selections by the true model and by the final fitted model. A large value in PCO and PFO indicates that most patients have received and will receive their optimal treatments determined by the true model, respectively.

The main differences between CIL & FIL and PCO & PFO are that for CIL and FIL, the penalties for non-matches depend on the individual loss; for PCO and PFO, all non-matches receive the same penalty. Small values of CIL and FIL, and large values of PCO and PFO are desirable.

5. Simulation studies

We performed a series of simulations to evaluate our design and compared it with the adaptive signature design and the adaptive enrichment design. For all the designs, we implemented the cross-validated procedures to search for final sensitive subsets from the entire patient population. To make ASD and AED comparable to SEDAR, all the designs used the Bayesian decision rules described above at final analyses.

We assume a two-treatment clinical trial designed to compare a control treatment (C) with an experimental treatment (E). We include situations where the experimental treatment is effective only in a sensitive subset. Let T_i denote the treatment indicator which

$$T_i = \begin{cases} 0.5 & \text{Treatment} = E \\ -0.5 & \text{Treatment} = C \end{cases}, \quad i = 1, \ldots, n. \tag{7}$$

5.1. Simulation setting

The total sample size N is set at 300; the size of the adaptive randomization blocks N_{bar} is set at 50. We choose the size N_0 of the initial block, for which equal randomization is used, to be $N_0 = 100$. Hence, we have a total of 5 blocks of sizes 100, 50, 50, 50 and 50. The tuning parameter for Bayesian adaptive randomization w is set at 0.5. The minimal clinical meaningful difference delta for sensitive patient status qualification is set at 0.1. The Type I error rate α_1 for testing treatment effect for the selected mixture of patients is controlled at 0.04, and the Type I error rate α_2 for testing treatment effect in the sensitive subset is controlled at 0.01.

We generate the ith patient's response from the following logistic regression model with ten biomarkers:

$$logit(r_i) = \beta_0 + \beta_1 T_i + \sum_{k=1}^{10} \gamma_k X_{ki} + \sum_{k=1}^{10} \eta_k T_i X_{ki}, \quad i = 1, \ldots, 300;\ k = 1, \ldots, 10. \tag{8}$$

where X_{ki} is the kth biomarker/signature for patient i. Each biomarker X_{ki} is assumed to follow a normal/multivariate normal distribution with mean of 0.5 and standard deviation of 1.0.

For the Bayesian adaptive randomization, we assume the prior is

$$\beta_0, \beta_1, \gamma_k, \eta_k \sim Normal(\mu = 0, \sigma^2 = 100).$$

At the final analysis, for each treatment arm, we use a non-informative prior $Beta(1, 1)$.

We consider three categories of biomarker distributions. For each type, we include four scenarios and a null scenario to compare the performance of SEDAR with ASD and AED (Table 2). The biomarker or biomarker-treatment interaction effects not presented in the table are assumed to be zero.

For the first category, we assume all biomarkers are independently distributed following a normal distribution with mean 0.5 and variance 1.0. For scenarios 1 and 2, we

Table 2. Simulation scenarios.

		β_0	β_1	γ_1	η_1	γ_6	η_6
x_1 to x_{10} are continuous with $\rho = 0.00$	Scenario 1	−0.8	0.0	0.0	0.0	0.0	0.0
	Scenario 2	−0.8	0.0	0.8	0.8	0.0	0.0
	Scenario 3	−0.8	0.0	0.3	1.1	0.0	0.0
	Scenario 4	−0.8	0.0	0.3	0.3	0.7	0.7
	Scenario 5	−0.8	0.0	0.4	0.4	1.1	1.1
x_1 to x_5 are continuous with $\rho = 0.25$	Scenario 6	−0.8	0.0	0.0	0.0	0.0	0.0
x_6 to x_{10} are continuous with $\rho = 0.00$	Scenario 7	−0.8	0.0	0.8	0.8	0.0	0.0
	Scenario 8	−0.8	0.0	0.3	1.1	0.0	0.0
	Scenario 9	−0.8	0.0	0.3	0.3	0.7	0.7
	Scenario 10	−0.8	0.0	0.4	0.4	1.1	1.1
x_1 to x_5 are continuous with $\rho = 0.75$	Scenario 11	−0.8	0.0	0.0	0.0	0.0	0.0
x_6 to x_{10} are continuous with $\rho = 0.25$	Scenario 12	−0.8	0.0	0.8	0.8	0.0	0.0
	Scenario 13	−0.8	0.0	0.3	1.1	0.0	0.0
	Scenario 14	−0.8	0.0	0.3	0.3	0.7	0.7
	Scenario 15	−0.8	0.0	0.4	0.4	1.1	1.1

consider only X_1 and the corresponding treatment interaction affect the response with different magnitude. For scenarios 3 and 4, we assume X_1 and X_6, and their interactions with treatment will affect the patient response with different scales.

For the second category, we assume the first five biomarkers have a multivariate normal distribution $X \sim NVM(0.5, \Sigma)$ with covariance matrix which has 1 for the variance and 0.25 for ρ. The other five variables are still independently distributed with a normal distribution with mean 0.5 and variance 1.0. For scenarios 7 and 8, we consider only X_1 and the corresponding treatment interaction affect the response. For scenarios 9 and 10, we assume one biomarker from the first five correlated biomarkers, X_1 and its interaction with treatment and another biomarker from the independent biomarker group, X_6 and its interaction with treatment, affect the response.

For the third category, we assume the first five biomarkers have a multivariate normal distribution with mean 0.5 and covariance matrix with $\rho = 0.75$, and the remaining five biomarkers have a multivariate normal distribution with mean 0.5 and covariance matrix with $\rho = 0.25$. The simulation scenarios are similar to the other two categories.

All the categories and scenarios are listed in Table 2. The null scenarios for each category numbers 1, 6 and 11, respectively. For each scenario, we generated 5000 simulations.

5.2. Simulation results

We compare the results obtained using our proposed design with those obtained using ASD and AED. The results for the estimated power are presented in Table 3. In the null scenarios, the Type I error rates for the treatment effect in the selected mixture patient group and the sensitive subgroup are controlled at 0.04 and 0.01, respectively.

For the rest of the scenarios, ASD always has the lowest power, while AED or SEDAR has the highest power. In Scenario 3, 8, 13, AED has a higher power than SEDAR. For example, in Scenario 3, the power in the patient mixture for ASD, AED and SEDAR are 0.744, 0.987 and 0.968, respectively. In the sensitive subgroup, the power is 0.721 for ASD,

Table 3. Comparison of three designs.

SCEN	Parameters	Design	Mixture power	Sens power	ORR	CIL	FIL	PCO	PFO
1	NULL	ASD	0.010	0.010	0.010	0.310	0.000	1.000	1.000
		AED	0.037	0.010	0.011	0.310	0.000	1.000	1.000
		SEDAR	0.038	0.011	0.008	0.310	0.000	1.000	1.000
2	$\gamma_1 = \eta_1 = 0.8$	ASD	0.510	0.300	0.410	0.073	0.031	0.501	0.715
		AED	0.815	0.560	0.461	0.088	0.014	0.500	0.902
		SEDAR	0.888	0.577	0.466	0.057	0.033	0.606	0.695
3	$\gamma_1 = 0.3$	ASD	0.744	0.721	0.353	0.107	0.023	0.500	0.766
	$\eta_1 = 1.1$	AED	0.987	0.920	0.374	0.127	0.009	0.500	0.909
		SEDAR	0.968	0.823	0.414	0.072	0.024	0.637	0.751
4	$\gamma_1 = \eta_1 = 0.3$	ASD	0.635	0.348	0.430	0.077	0.032	0.500	0.723
	$\gamma_6 = \eta_6 = 0.7$	AED	0.854	0.590	0.472	0.090	0.014	0.500	0.911
		SEDAR	0.911	0.614	0.483	0.059	0.034	0.610	0.701
5	$\gamma_1 = \eta_1 = 0.4$	ASD	0.794	0.602	0.478	0.097	0.028	0.500	0.771
	$\gamma_6 = \eta_6 = 1.1$	AED	0.964	0.833	0.537	0.111	0.012	0.500	0.927
		SEDAR	0.993	0.892	0.557	0.067	0.030	0.639	0.760
6	NULL	ASD	0.040	0.008	0.310	0.000	0.000	1.000	1.000
		AED	0.040	0.008	0.310	0.000	0.000	1.000	1.000
		SEDAR	0.040	0.010	0.310	0.000	0.000	1.000	1.000
7	$\gamma_1 = \eta_1 = 0.8$	ASD	0.523	0.327	0.410	0.073	0.031	0.501	0.718
		AED	0.807	0.587	0.463	0.088	0.013	0.499	0.908
		SEDAR	0.894	0.594	0.468	0.057	0.033	0.607	0.697
8	$\gamma_1 = 0.3$	ASD	0.744	0.693	0.353	0.107	0.023	0.500	0.767
	$\eta_1 = 1.1$	AED	0.990	0.931	0.377	0.128	0.009	0.500	0.917
		SEDAR	0.974	0.842	0.415	0.073	0.024	0.636	0.752
9	$\gamma_1 = \eta_1 = 0.3$	ASD	0.653	0.362	0.430	0.077	0.032	0.500	0.725
	$\gamma_6 = \eta_6 = 0.7$	AED	0.855	0.587	0.471	0.090	0.014	0.500	0.911
		SEDAR	0.917	0.638	0.482	0.058	0.034	0.612	0.706
10	$\gamma_1 = \eta_1 = 0.4$	ASD	0.808	0.615	0.478	0.097	0.028	0.501	0.771
	$\gamma_6 = \eta_6 = 1.1$	AED	0.971	0.826	0.537	0.111	0.013	0.500	0.926
		SEDAR	0.991	0.895	0.555	0.068	0.030	0.637	0.756
11	NULL	ASD	0.037	0.010	0.310	0.000	0.000	1.000	1.000
		AED	0.039	0.010	0.310	0.000	0.000	1.000	1.000
		SEDAR	0.038	0.009	0.310	0.000	0.000	1.000	1.000
12	$\gamma_1 = \eta_1 = 0.8$	ASD	0.514	0.329	0.412	0.073	0.029	0.500	0.732
		AED	0.851	0.625	0.468	0.090	0.010	0.500	0.922
		SEDAR	0.898	0.620	0.473	0.058	0.031	0.608	0.708
13	$\gamma_1 = 0.3$	ASD	0.737	0.729	0.353	0.107	0.022	0.501	0.785
	$\eta_1 = 1.1$	AED	0.992	0.945	0.379	0.131	0.007	0.500	0.938
		SEDAR	0.980	0.885	0.419	0.074	0.024	0.640	0.768
14	$\gamma_1 = \eta_1 = 0.3$	ASD	0.637	0.351	0.428	0.077	0.032	0.500	0.723
	$\gamma_6 = \eta_6 = 0.7$	AED	0.852	0.585	0.472	0.090	0.013	0.500	0.914
		SEDAR	0.909	0.631	0.485	0.059	0.035	0.610	0.696
15	$\gamma_1 = \eta_1 = 0.4$	ASD	0.798	0.606	0.478	0.097	0.029	0.501	0.770
	$\gamma_6 = \eta_6 = 1.1$	AED	0.972	0.849	0.539	0.112	0.012	0.500	0.928
		SEDAR	0.991	0.880	0.557	0.068	0.031	0.639	0.753

0.920 for AED and 0.823 for SEDAR. For the other scenarios, SEDAR achieves the highest power in both the mixture of patients and sensitive subset. For scenario 7, the power in the mixture of patients is 0.523 for ASD, 0.807 for AED and 0.894 for SEDAR.

The results of the four additional trial performance evaluation criteria we proposed in Section 4 are presented in Table 3. While both ASD and AED adopt an equal randomization approach when assigning patients to treatment arm, resulting in equal numbers of patients per treatment arm, SEDAR assigns more patients to the superior treatment based on baseline biomarker profiles and data accumulation during the trial. As a result, SEDAR always achieves a higher objective response rate (ORR) than ASD and AED. For example, in scenario 15, 55.7% of patients respond to the assigned treatment for SEDAR compared to 47.8% for ASD and 53.9% for AED. AED results in a higher proportion of patients who respond to the assigned treatment than ASD. This demonstrates that the enrichment strategy allows more trial participants to achieve responses than the all-comers designs.

CIL and FIL measure the loss for trial participants and future patients. ASD and AED have similar values of CIL, while SEDAR has much smaller values than the other designs. Using scenario 5 as an example, CIL for SEDAR is 0.067 compared to 0.097 for ASD and 0.111 for AED. This difference in CIL shows that SEDAR results in more trial participants receiving optimal treatments. For FIL, the results are similar for ASD and SEDAR, but AED has the smallest values in all scenarios. This shows that a complete enrichment strategy may allow more future patients to receive their optimal treatments.

SEDAR also has higher PCO values than the other designs. That is, SEDAR results in a higher proportion of enrolled patients receiving their optimal treatments as defined by the true model. Similar to the relative performance among the three designs, SEDAR and ASD show similar results for PFO, which refers to the probability for a future patient receiving the personalized optimal treatment if the treatment recommendation from the trial were followed.

The results for true and the estimated treatment effects are presented in the supplemental materials. Tables 1S and 2S show the results for the selected mixtures of sensitive and nonsensitive patients. Tables 4S and 5S show the results for the sensitive patient subset. In these tables, the estimated treatment effects and true effects match well. Table 3S and 6S show the proportions of truly sensitive patients among the selected mixtures, and the selected sensitive patients by the final fitted model. According to the results, AED trials have the highest proportion of truly sensitive patients, following by SEDAR and ASD, demonstrating that enrichment strategies include more patients who are likely to respond to the treatment than all comers designs, one of the key reasons for considering enrichment designs.

More research is needed for estimating the causal effects of experimental treatment vs. control. This is a challenging problem for many enrichment designs, especially when the subsets for defining enrichment strategy can not be pre-specified due to lack of data and knowledge on which biomarkers affect treatment effects. This is the situation we consider in this article, in which the enrichment design itself serves as a search engine for sensitive subsets.

6. Example

To illustrate our approach, consider the design of a trial in non-small-cell lung cancer (NSCLC) comparing two treatment arms, one using a tyrosine kinase inhibitor (gefitinib) and the other using standard chemotherapy (carboplatin plus paclitaxel), similar to the Iressa Pan-Asia Study (IPASS) [28]. The single binary biomarker in this case is the EGFR

Table 4. IPASS example: biomarker settings.

Biomarkers	Gefitinib	Chemotherapy
EGFR status		
EGFR +	0.50	0.50
EGFR −	0.50	0.50
Age		
$<$ 65 years	0.72	0.76
$\geq$ 65 years	0.28	0.24
Gender		
Male	0.20	0.21
Female	0.80	0.79
Smoking history		
Non-smoker	0.938	0.936
Former light smoker	0.062	0.064
WHO performance status		
0 or 1	0.90	0.89
2	0.10	0.11

mutation status, either positive or negative, with a prevalence of EGFR positive patients of approximately 50% in Asian countries and approximately 20% in North America [27]. For the purpose of illustrating the application in the setting of this paper (binary response assessed soon after the start of treatment), we consider the endpoint to be objective response rate. Unfortunately, since the assessment of EGFR mutation status was not an eligibility requirement, only 36% of the patients (437 of 1217) had a known status. To conduct simulations using the IPASS study as a guide, we assume the prevalence of positive EGFR is 0.50 (commonly observed in an Asian population) and the response rates in the various treatment arms and EGFR mutation status groups are as follows: 0.45 in the chemotherapy arm for EGFR positive; 0.70 in the gefitinib arm for EGFR positive; 0.15 in the chemotherapy arm for EGFR negative; and 0.10 in the gefitinib arm for EGFR negative patients.

Although this example is presented as an illustration of the setting considered here, the final endpoints in such a study would ordinarily be progression-free survival (PFS) or overall survival (OS), not the response rates. In the initial results reported for the IPASS study, there was improved PFS for gefitinib in the EGFR mutation-positive group but worse PFS for gefitinib in the EGFR mutation-negative group. These results are suggested in the early results but not definitively so except perhaps in the EGFR negative group because of the dramatic early difference.

To apply our design to the IPASS study, we assume a total sample size of 440; an initial block size of 176 patients, and subsequent adaptive randomization block size of 88 patients. The overall Type I error rate is controlled at 0.05 with 0.04 for testing the treatment effect in the mixture of sensitive and non-sensitive patients and 0.01 for testing in the sensitive subset only. The clinical meaningful difference for enrolling sensitive patients is set at 0.1.

Following the approach used in the IPASS paper for progression-free survival [28], we use the following covariates in a logistic regression model for ORR, all as binary covariates: Age ($<$ 65 years vs. $\geq$ 65 years), Sex (Male, Female), Smoking history (non-smoker vs. former light smoker), WHO performance status (0 or 1 vs. 2). We assume only the EGFR mutation status and its interaction have a treatment effect. In the true model, no treatment main effect is included. The prevalence of each biomarker stratum is presented in Table 4.

Table 5. Simulation results: IPASS example.

Design	Mixture Power	Sens Power	ORR	CIL	FIL	PCO	PFO
ASD	0.738	0.946	0.361	0.096	0.089	0.501	0.605
AED	0.995	0.993	0.442	0.121	0.035	0.449	0.838
SEDAR	0.999	0.982	0.446	0.063	0.051	0.662	0.768

For the given design parameters, the results are presented in Table 5. Generally, the results confirm the conclusion drawn from the previous simulation studies. The power of the mixture population is 0.738 for ASD; 0.995 for AED; and 0.999 for SEDAR, and of the sensitive subset is 0.946 for ASD, 0.993 for AED and 0.982 for SEDAR. The objective response rates among all trial participants are 0.361 by ASD; 0.442 by AED; and 0.446 by SEDAR. In addition, SEDAR also has the lowest CIL index and high values for PCO. Thus, our proposed design can assign more patients to receive their optimal treatments.

7. Discussion

In this article, we have proposed a signature enrichment design with adaptive randomization for cancer clinical trials. Our proposed design builds on the advantages of the adaptive signature design and the adaptive enrichment design. Assuming that patient responses can be observed shortly after receiving treatment, our design adopts an adaptive enrichment approach by developing a sensitive patient classifier to screen sensitive and non-sensitive patients for enrollment. Once a patient enters the trial, we use response-adaptive randomization to adjust treatment allocation probabilities in a Bayesian framework. At the final analysis, we use an adaptation of the adaptive signature designs. For testing the treatment effect in the subset of sensitive patients, this design uses a cross-validated procedure to efficiently identify sensitive patients from all the trial participants.

To address the *personalized* aspect of precision medicine, we also proposed four trial performance evaluation criteria in addition to the traditional operating characteristics. The criteria include CIL, FIL, PCO and PFO for both currently enrolled patients and future patients. These criteria numerically present how well the trial design will offer patients their own personalized optimal treatment.

Our results have shown that ASD always has the lowest power, while SEDAR has the highest power in most scenarios. The main difference between AED or SEDAR and ASD is that both AED and SEDAR are ‘enrichment’ designs, enrolling more ‘sensitive’ patients than ASD, which enrolls all patients without selection.

To help clarify the relative performance of power between AED and SEDAR, denote by $r|E$ the response rate among all patients in a trial who receive the experimental treatment (E), and $r|C$ the response rate among those who receive the control treatment (C). The main reason for the better power of SEDAR versus AED in most of our simulation scenarios is that $r|E - r|C$ is larger in the SEDAR trials than in the AED trials. This is caused by the covariate-based outcome-adaptive randomization adopted by SEDAR. A toy example is given below.

Suppose the population consists of two groups of patients: 1/2 sensitive and 1/2 non-sensitive. For sensitive patients, the response rates by E and C are 0.6 and 0.3, respectively. For non-sensitive patients, the response rate is 0.3 regardless of treatment. Assume the total sample size is 300. By ASD, due to its non-selective enrollment and equal randomization, it is easy to see that it results in $r|E - r|C = 0.15$. Among the patients who received E, the ratio between sensitive and nonsensitive patients is 1:1. Suppose that by the AED design, 80 non-sensitive patients and 220 sensitive patients are enrolled. Due to the equal randomization, among each group (sensitive or non-sensitive), 50% patients have received E, and 50% received C. Then in this setting, $r|E = 0.52$, $r|C = 0.30$, thus $r|E - r|C = 0.22$. Among the patients who received E, the ratio between sensitive and nonsensitive patients is 2.75:1.

For SEDAR, assume 100 non-sensitive and 200 sensitive patients are enrolled. By SEDAR, non-sensitive patients are equally randomized between E and C, whereas sensitive patients are more likely to be assigned to E. Assume 80% of sensitive patients have received E, and 20% received C. In this setting, $r|E \approx 0.53$, $r|C = 0.30$, thus $r|E - r|C \approx 0.23$. Among the patients who received E, the ratio between sensitive and nonsensitive patients is 3.2:1. The effect size 0.23 may appear to be inflated by over-sample the sensitive patients for arm E. However, if the arm C also had a ratio of 3.2:1 between sensitive and nonsensitive patients, the effect size would still be 0.23 since that sensitive and nonsensitive patients have the same response rate by the control treatment. In this sense, the effect size 0.23 is still valid, but we must make it clear that this is for a special population with mixture ratio 3.2:1 between sensitive and nonsensitive patients.

SEDAR results in a higher value of $r|E - r|C$ than ASD or AED, thus higher power. Considering the fact these three designs end up different mixtures of sensitive and non-sensitive patients, this may not sound like a fair comparison of power between them in the traditional sense. However, this higher 'power' by SEDAR does indicate that its better ability to end up with a mixture of patients that best reflect the benefit of E over C, thus leading to better chance of new drug discovery.

This brings in other important questions. How to estimate the 'treatment effects' for the AED and SEDAR trials? Do the above results show that SEDAR introduces bias or overestimate the benefit of E over C? This depends on how we interpret the results from the trial. In the above example, since the treatment effects of C are the same in sensitive and nonsensitive patients, the ratio of these two groups of patients has no impact on $r|C$. Therefore, when this ratio is 3.2:1, we have $r|E - r|C \approx 0.23$. Based on this interpretation, there is not a problem of bias. However, more careful considerations are needed for estimating and reporting treatment effects for enrichment trials, as described below.

All enrichment designs need extra work to estimate an overall treatment effect due to their selective enrollment. They all have to report subset-specific treatment effects. For example, those enrichment designs utilizing pre-specified categories, such as positive and negative subgroups of some specific biomarkers, usually report treatment effects for each category. Depending on whether and how many biomarker negative patients are enrolled, such an enrichment trial may or may not report the treatment effects for the biomarker negative subgroups. If biomarker negative patients are enrolled but undersampled, then special considerations are needed for the estimation of the overall treatment effect. Moreover, such pre-specified categories may not exist or well-defined before a trial is conducted. The effects of many targeted therapies cannot be well separated by the positive/negative status of biomarkers. Even the cut-off values for biomarker positive/negative status may not be well established before a trial is conducted. Our proposed design is proposed for such a situation. With multiple biomarkers, it may not be feasible to make inference for all the possible combination of biomarker positive/negaitve subgroups because some of them may have a small sample size. Therefore, we fit models to select sensitive patients. Then the treatment effect estimation become complicated in this situation. It requires additional work to estimate the overall treatment effect, which needs to account for the time-varying treatment selection procedure. This warrants future research.

For clinical trials that do not consider patients' biomarkers X or in which the biomarker is not predictive, the underlying assumption is that all patients are homogeneous and the patient heterogeneity is ignored. In this situation, it is true that, when compared with equal randomization, adaptive randomization will lose power with some exceptions (e.g. some scenarios with binary outcomes). However, in this paper, we account for patient heterogeneity and use regression models of response on X to predict future patients response rates to E or C. Although both AED and SEDAR use these results to select patients to enroll, only SEDAR uses these results to adjust patients' probability to receive E or C. The consequence is that, as seen from above, SEDAR trials have larger values of $r|E - r|C$, and thus higher power levels than AED.

The purpose of our proposal is not to compare three different designs with respect to power within a fixed patient population. Our purpose is to conduct a randomized phase II trial to identify subpopulations in which a new drug would work better than the standard treatment. In this context, the higher power performance for SEDAR indicates its better ability to identify such subpopulations. This means there is a higher chance to success using the proposed SEDAR design rather than ASD or AED. This makes SEDAR a useful design for pharmaceutical development. Our proposed design is not intended for phase III confirmatory trials; rather it is for phase II exploratory trials.

Another feature in our designs affects the power. In the simulations, for AED and SEDAR, we have used slightly different enrollment functions for non-sensitive patients. While AED gives a zero probability for non-sensitive patients to enroll, SEDAR gives a probability between 0.1 and 1 for them. This difference favors AED for power (i.e. should make AED have higher power than SEDAR). The final power is the combined effect of this enrollment function and the above effect on $r|E - r|C$ caused by the adaptive randomization.

Our results have also shown that adopting an enrichment strategy – oversampling the sensitive patients and undersampling the nonsensitive patients will offer benefits to trial participants. This is particularly important for late-stage cancer patients since many will need to enroll in clinical trials to have access to novel treatments. The enrichment strategy described here can be applied to other biomarker adaptive designs [8,13,25,38]. For example, the recently proposed enriched biomarker stratified design and the auxiliary-variable-enriched biomarker stratified design also show that applying an enrichment strategy can result in a more cost-efficient design in terms of power [40,41]. For both of these designs, the issue is how to choose the optimal enrichment proportion. Our proposed design uses a continuous loss function based on the response rate estimates for each treatment for a given patient using a classification algorithm. In addition to using an enrichment strategy to increase trial efficiency, our proposed design also utilizes Bayesian adaptive randomization approach to sequentially update treatment assignment probabilities based on patients' biomarker signatures measured at baseline. The simulation results imply that our design has more promising trial operating characteristics in most scenarios.

Although our design has distinct advantages, there are some caveats. The complexity of utilizing an enrichment strategy as well as a Bayesian adaptive randomization scheme, yields operational complexities. In addition, in common with other types of adaptive enrichment trials, there is a potential for the trial duration to increase and raise issues of interpretation of the final tests and treatment effect estimates arising from the adaptive enrichment process. Nevertheless, our proposed design allows more patients to receive optimal personalized treatments, thus yielding a higher overall response rate on the trial. It can identify therapies that are effective only in a sensitive subset, increase the chance to discover a successful new drug, and outperform the all-comers equal randomization trial designs in many scenarios.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the U.S. National Science Foundation under Grant Division of Mathematical Sciences (DMS) [grant number 1612965]; the National Cancer Institute of the National Institutes of Health [grant numbers U54CA096300, U01CA152958 and 5P50CA100632]. The research of Huang was partially supported by the Dr Mien-Chie Hung and Mrs Kinglan Hung Endowed Professorship.

References

[1] D.A. Berry, *Adaptive clinical trials:the promise and the caution*, J. Clin. Oncol. 29 (2011), pp. 606–609.

[2] D.A. Berry, R.S. Herbst, and E.H. Rubin, *Reports from the 2010 clinical and translational cancer research think tank meeting: design strategies for personalized therapy trials*, Clin. Cancer Res. 18 (2012), pp. 638–644.

[3] L. Breiman, *Random forest*, Mach. Learn. 45 (2001), pp. 5–32.

[4] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Chapman & Hall/CRC, Boca Raton, 1984.

[5] Y. Cheung, L. Inoue, J. Wathen, and P. Thall, *Continuous Bayesian adaptive randomization based on event times with covariates*, Stat. Med. 25 (2006), pp. 55–70.

[6] S. Dudoit, J. Fridlyand, and T.P. Speed, *Comparison of discrimination methods for the classification of tumors using gene expression data*, J. Am. Stat. Assoc. 97 (2002), pp. 77–87.

[7] B. Freidlin, W. Jiang, and R. Simon, *The cross-validated adaptive signature design*, Clin. Cancer Res. 16 (2010), pp. 691–698.

[8] B. Freidlin, L.M. McShane, and E.L. Korn, *Randomized clinical trials with biomarkers: design issues*, J. Natl. Cancer Inst. 102 (2010), pp. 152–160.

[9] B. Freidlin and R. Simon, *Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients*, Clin. Cancer Res. 11 (2005), pp. 7872–7878.

[10] P. Gallo, C. Chuang-Stein, V. Dragalin, B. Gaydos, M. Krams, and J. Pinheiro, *Adaptive design in clinical drug development – an executive summary of the phrma working group (with discussions)*, J. Biopharm. Stat. 16 (2006), pp. 275–283.

[11] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, 2nd ed., Chapman and Hall/CRC, Boca Raton, 2004.

[12] S.L. George, *Statistical issues in translational cancer research*, Clin. Cancer Res. 14 (2008), pp. 5954–5958.

[13] S.L. George, X. Wang, and H. Pang, eds., *Cancer Clinical Trials*, Chapman and Hall/CRC, New York, 2016.

[14] K.A. Gold, E.S. Kim, J.J. Lee, I.I. Wistuba, C.J. Farhangfar, and W.K. Hong, *The BATTLE to personalize lung cancer prevention through reverse migration*, Cancer Prev. Res. (Philadelphia, PA) 4 (2011), pp. 962–972.

[15] V. Grunwald and M. Hidalgo, *Developing inhibitors of the epidermal growth factor receptor for cancer treatment*, J. Natl. Cancer Inst. 95 (2003), pp. 851–867.

[16] E. Hade, D. Jarjoura, and L. Wei, *Sample size re-estimation in a breast cancer trial*, Clin. Trials 7 (2010), pp. 219–226.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, New York, NY, 2009.

[18] B.P. Hobbs, B.P. Carlin, and D.J. Sargent, *Adaptive adjustment of the randomization ratio using historical control data*, Clin. Trials (London, England) 10 (2013), pp. 430–440.

[19] F. Hu and W. Rosenberger, *The Theory of Response-Adaptive Randomization in Clinical Trials*, Wiley, Hoboken, NJ, 2006.

[20] E.L. Korn and B. Freidlin, *Adaptive clinical trials:advantages and disadvantages of various adaptive design elements*, JNCI: J. Natl. Cancer Inst. 109 (2017), p. djx013.

[21] J. Kruppa, Y. Liu, G. Biau, M. Kohler, I.R. König, J.D. Malley, and A. Ziegler, *Probability estimation with machine learning methods for dichotomous and multicategory outcome:applications*, Biometr. J. 56 (2014), pp. 564–583.

[22] J. Lee, X. Gu, and S. Liu, *Bayesian adaptive randomization designs for targeted agent development*, Clin. Trials 7 (2010), pp. 584–596.

[23] Q. Liu, M. Proschan, and G. Pledger, *A unified theory of two-stage adaptive designs*, J. Am. Stat. Assoc. 97 (2002), pp. 1034–1041.

[24] S. Mandrekar and D. Sargent, *Randomized phase II trials:time for a new era in clinical trial design*, J. Thorac. Oncol. 5 (2010), pp. 932–934.

[25] S. Matsui, M. Buyse, and R. Simon, eds., *Design and Analysis of Clinical Trials for Predictive Medicine*, Chapman and Hall/CRC, New York, 2015.

[26] C. Meacham and S. Morrison, *Tumor heterogeneity and cancer cell plasticity*, Nature 501 (2013), pp. 328–337.

[27] A. Midha, S. Dearden, and R. McCormack, *EGFR mutation incidence in non-small-cell lung cancer of adenocarcinoma histology: a systematic review and global map by ethnicity (mutMAPII)*, Am. J. Cancer Res. 5 (2015), pp. 2892–2911.

[28] T. Mok, Y. Wu, S. Thongprasert, C. Yang, D. Chu, N. Saijo, P. Sunpaweravong, B. Han, B. Margono, Y. Ichinose, Y. Nishiwaki, Y. Ohe, J. Yang, B. Chewaskulyong, H. Jiang, E. Duffield, C. Watkins, A. Armour, and M. Fukuoka, *Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma*, New Engl. J. Med. 361 (2009), pp. 947–957.

[29] H. Moon, H. Ahn, R. Kodell, S. Baek, C.J. Lin, and J.J. Chen, *Ensemble methods for classification of patients for personalized medicine with high-dimensional data*, Artif. Intell. Med. 41 (2007), pp. 197–207.

[30] J. Ning and X. Huang, *Response-adaptive randomization for clinical trials with adjustment for covariate imbalance*, Stat. Med. 29 (2010), pp. 1761–1768.

[31] Q. Shi, S. Mandrekar, and D. Sargent, *Predictive biomarkers in colorectal cancer: usage, validation, and design in clinical trials*, Scand. J. Gastroenterol. 47 (2012), pp. 356–362.

[32] R. Simon, *Development and validation of biomarker classifiers for treatment selection*, J. Stat. Plan. Inference 138 (2008), pp. 308–320.

[33] R. Simon, *Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology*, Pers. Med. 7 (2010), pp. 33–47.

[34] R. Simon, E. Korn, L. McShane, M. Radmacher, G. Wright, and Y. Zhao, *Design and Analysis of DNA Microarray Investigations*, Springer, New York, NY, 2004.

[35] N. Simon and R. Simon, *Adaptive enrichment designs for clinical trials*, Biostatistics 14 (2013), pp. 613–625.

[36] P. Thall and J. Wathen, *Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments*, Stat. Med. 24 (2005), pp. 1947–1964.

[37] P. Thall and J. Wathen, *Practical Bayesian adaptive randomisation in clinical trials*, Eur. J. Cancer (Oxford, England: 1990) 43 (2007), pp. 859–866.

[38] R. Uozumi and C. Hamada, *Interim decision-making strategies in adaptive designs for population selection using time-to-event endpoints*, J. Biopharm. Stat. 27 (2017), pp. 84–100.

[39] V. Vladimir, *The Nature of Statistical Learning Theory*, 2nd ed., Springer, New York, NY, 1999.

[40] T. Wang, X. Wang, H. Zhou, J. Cai, and S.L. George, *Auxiliary-variable-enriched biomarker-stratified design*, Stat. Med. 37 (2018), pp. 4610–4635.

[41] X. Wang, J. Zhou, T. Wang, and S.L. George, *On enrichment strategies for biomarker stratified clinical trials*, J. Biopharm. Stat. 28 (2018), pp. 292–308.

[42] J. Wathen and P. Thall, *A simulation study of outcome adaptive randomization in multi-arm clinical trials*, Clin. Trials 14 (2017), pp. 432–440.

[43] Y. Yuan, X. Huang, and S. Liu, *A Bayesian response-adaptive covariate-balanced randomization design with application to a leukemia clinical trial*, Stat. Med. 30 (2011), pp. 1218–1229.